Assembling the marine metagenome, one cell at a time

Tanja Woyke[1], Gary Xie[3], Alex Copeland[1], José M. González[4], Cliff Han[3], Hajnalka Kiss[3], Jimmy Saw[3], Pavel Senin[3], Chi Yang[3], Jan-Fang Cheng[1], Michael E. Sieracki[2] and Ramunas Stepanauskas[2]

[1] DOE Joint Genome Institute, Walnut Creek, California, [2] Bigelow Laboratory for Ocean Sciences, West Boothbay Harbor, Maine, [3] Los Alamos National Laboratory, Los Alamos, New Mexico
[4] Department of Microbiology, University of La Laguna, La Laguna, Tenerife, Spain

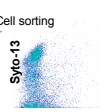
Abstract

Determining the genetic makeup of predominant microbial taxa with specific metabolic capabilities remains one the major challenges in microbial ecology and bioprospecting, due to the limitations of current cell culturing and metagenomic methods. The complexity of microbial communities and intraspecies variations poses obstacles for the assembly of individual genomes from metagenomic shotgun libraries. Here we report the use of single cell genomics to access the genome of two uncultured proteorhodopsin-encoding flavobacteria from Gulf of Maine. High throughput fluorescence-activated sorting of single cells, whole genome amplification via multiple displacement amplification and PCR-screening enabled shotgun sequencing of these single amplified genomes (SAGs) yielding 1.9 Mbp (17 contigs) and 1.5 Mbp (21 contigs) draft genomes for the two flavobacteria. In contrast to cultured strains, the two uncultured flavobacteria genomes were excellent Global Ocean Sampling (GOS) metagenome fragment recruiters, demonstrating their numerical significance in the ocean. Annotation revealed genome streamlining and diversified energy sources of the two uncultured microorganisms, including biopolymer degradation, proteorhodopsin photometabolism, and H2 oxidation. These features may be indicative for specific adaptations to the marine environment and for the absence of related microorganisms in cultures.

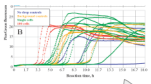
Background & Methods

Costal water (Maine)

Cell sorting



Real-time multiple displacement amplification (MDA) of single cells



16S rRNA gene and proteorhodopsin gene PCR of single cells

Stepanauskas, PNAS 2007.

Sequencing

	MS024-2A	MS024-3C
454 data	73 Mb	90 Mb
3Kb Sanger	4.58 Mb	5 Mb
8Kb Sanger	12.87 Mb	17.67 Mb

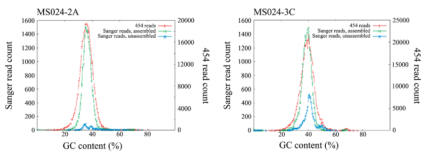
Assembly features

General features of the flavobacterial SAG assemblies.

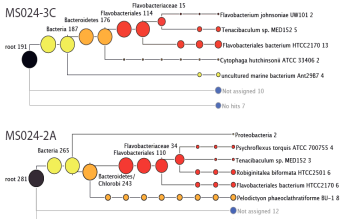
	MS024-2A	MS024-3C
Assembly statistics		
Assembly size [Mbp]	1.9	1.5
Estimated genome size [Mbp]	2.1	1.9
Estimated genome recovery [%]	91	78
Number of contigs	17	21
Largest contig [bp]	684,032	549,383
GC content [%]	36	39
Mean read depth (± sd)	56 (± 63)	83 (± 110)
454 reads	47	68
Sanger reads	9	14.3
Gene predictions		
Total genes	1,815	1,413
Protein coding genes	1,780	1,388
with function prediction	1,337	1,060
w/o function prediction	443	328
Number of rRNA operons	2	1
Number of tRNA genes	33	24

The sequence data of the SAGs was Phrap assembled, followed by primer walking on shotgun clones, and PCR/adapter PCR on the diluted MDA products.

Data QC

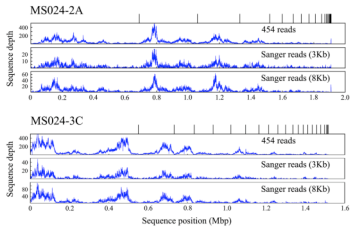


GC contents histogram of the unassembled and assembled Sanger and pyrosequence reads for the two SAG exhibits a tight unimodal distribution.

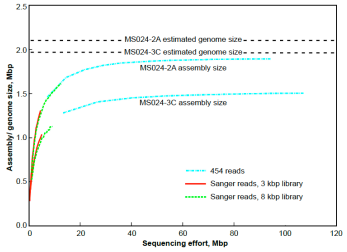


Taxonomic assignment analysis. The taxonomic contents of the blastx output for the unassembled reads of the Flavobacteria sp. MS024-2A was estimated and visualized using the Metagenome Analyzer (MEGAN) (Huson, Genome Res 2007).

The MDA bias



MDA bias as evaluated by sequence depth distribution. The contigs for the SAG are aligned by length and contig breaks are indicated by the tic marks along the top. The mean sequence depth is 56 (± 63) for MS024-2A and 83 (±110) for MS024-3C.



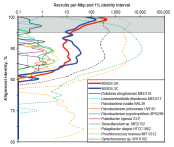
Genome coverage as function of the genome sequencing effort for the flavobacterial SAG. The curve displays near-saturation indicating that additional sequencing would mostly result in repeated sampling of the over-amplified genomic regions, not targeting the yet missing part of the genome.

Chimeras

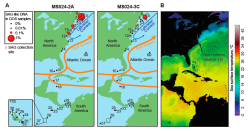
Chimeric rearrangements in the SAG DNA.

	MS024-2A		MS024-3C	
	chimeric reads/ clones (%)	overall chimerism	chimeric reads/ clones (%)	overall chimerism
Read-based chimerism				
34Kb library reads (untreated MDA DNA)	1.9	1 chimer/ 28 Kbp	NA	NA
34Kb library reads (S1 treated MDA DNA)	2.0	1 chimer/ 26 Kbp	1.9	1 chimer/ 33 Kbp
8Kb library reads (S1 treated MDA DNA)	2.1	1 chimer/ 30 Kbp	1.6	1 chimer/ 40 Kbp
454 reads (S1 treated MDA DNA)		1 chimer/ 19 Kbp		1 chimer/ 29Kbp
Average (all reads)		1 chimer/ 21 Kbp		1 chimer/ 27Kbp
Clone-based chimerism				
34Kb clones (untreated MDA DNA)	14.0	1 chimer/ 20 Kbp	NA	NA
paired reads facing into the same direction	8.6			
paired reads facing away from insert	3.8			
paired reads outside the insert size range	0.2			
paired reads in different contigs	0.8			
paired reads contained in each other	1.1			
34Kb clones (S1 treated MDA DNA)	**16.8**	**1 chimer/ 15 Kbp**	**16.4**	**1 chimer/ 20 Kbp**
paired reads facing into the same direction	7.8		9.6	
paired reads facing away from insert	3.0		0.9	
paired reads outside the insert size range	3.2		4.6	
paired reads in different contigs	2.0		0.8	
paired reads contained in each other	0.8		0.4	
8Kb clones (S1 treated MDA DNA)	26.4		29.8	1 chimer/ 27Kbp
paired reads facing into the same direction	22.6	1 chimer/ 17Kbp	17.4	
paired reads facing away from insert	2.6		2.7	
paired reads outside the insert size range	7.4		3.9	
paired reads in different contigs	3.3		3.8	
paired reads contained in each other	0.6		1.8	
Average (all clones)		**1 chimer/ 17Kbp**		**1 chimer/ 23Kbp**

Fragment recruitment

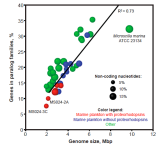


Global Ocean Sampling (GOS) (Rusch, PLoS Biol 2007) metagenome fragment recruitment by the SAGs MS024-2A and MS024-3C, the currently sequenced marine Flavobacteria isolate genomes, the non-marine *F. johnsoniae*, and the three best GOS fragment recruiters *Pelagibacter*, *Prochlorococcus* and *Synechococcus*.



Geographic distribution of the GOS (Rusch, PLoS Biol 2007) metagenome fragments with >95% nucleotide identity to MS024-2A and MS024-3C.

Genome Streamlining



Genome streamlining in MS024-2A and MS024-3C was evidenced by small genome sizes, low fraction of genes in paralog families, and low fraction of non-coding bases. Included are all available genomes of the Bacteroidetes/Chlorobi group. The number of genes in paralog families was estimated using the BLASTCLUST tool from the NCBI BLAST software (>30% sequence similarity, across >50% of their length and E<10-6).

Conclusion

Using the single cell approach, we demonstrate how a combination of single cell FACS and amplification via MDA can be used to access the genomes of uncultured environmental microorganisms, representative of their given environment.

Acknowledgements

We would like to thank PGF for the sequencing efforts and Lynne Goodwin (Los Alamos National Laboratory) for the coordination of the efforts involved in this project.

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.